計量テキスト分析およびKH Coderを用いた論文の執筆・査読チェックポイント
――明示すべき点と分析結果の記述を中心に――

中村　康則[i]，周　景龍[ii]，樋口　耕一[iii]

計量テキスト分析とは，文章データを統計的に処理し，内容分析を行なう方法である。この計量テキスト分析のためのツールとして筆者らが開発したKH Coderは，社会科学・人文科学の分野を中心に，8,000件を超えるさまざまな学術研究で活用されている。本稿では，KH Coderによる計量テキスト分析の結果を用いて，適切に論文を執筆するためのチェックポイントを整理する。そして各チェックポイントについて，なるべく実際の研究事例から，優れた実践のあり方を示す。これらのチェックポイントは，計量テキスト分析を用いた論文を査読する際にも有用であろう。

キーワード：計量テキスト分析，テキストマイニング，KH Coder，査読，研究事例レビュー

はじめに

執筆と査読のチェックポイント

計量テキスト分析とは，たとえば新聞記事やアンケート自由記述，インタビュー記録のようなテキスト型データを計量的に分析する方法である（川端 2021; 樋口 2004, 2020）。計量テキスト分析のためのソフトウェアとして筆者らが開発したKH Coderは[1)]，社会科学・人文科学の分野を中心に8,000件を超える学術研究で活用されている。学術研究のほかにも行政，医療・看護，マーケティングといったさまざまな領域で利用例が見られる。さらにKH Coderは日本語・英語を始め，中国語・スペイン語を含む多くの言語データの分析に対応しており，海外でも一部利用されつつある。

このKH Coderを使って分析を行なう方法や，KH Coderの操作方法については，すでに樋口・中村・周（2022）に主要な部分をまとめている。KH Coderは多くの機能を備えているが，「これだけ押さえておけば問題なく分析を行なえるだろう」という機能を厳選して樋口・中村・周（2022）で紹介している。それ以外の機能について知りたい場合や，各機能についてより詳しくは樋口（2020）を参照されたい。

このようにKH Coderを操作して分析を行なう方法についてはいくつかの文献で示してきた。しかし，分析結果を用いた適切な論文の書き方については，これまでのところ十分に提示できていない。筆者ら自身による応用研究を発表し，事例として参考に供してはいるものの，ガイドラインとしてより使いや

i　京都大学大学院医学研究科非常勤講師・株式会社SCREENアドバンストシステムソリューションズ執行役員
ii　株式会社SCREENアドバンストシステムソリューションズ開発部開発担当課長
iii　立命館大学産業社会学部教授

すいものがあった方が望ましいだろう。

そこで本稿ではKH Coderによる計量テキスト分析の結果を，適切に論文に用いるために留意すべき点を整理する。そして各チェックポイントについて，なるべく実際の研究事例から，優れた実践のあり方を示す。これらのチェックポイントは，論文を書くときだけでなく，計量テキスト分析を用いた論文を査読する際にも有用であろう。

なおデータの統計分析になじみが深い研究分野では，これらのチェックポイントは当然のことと感じられるかもしれない。しかし近年では，必ずしもそうではない分野でもKH Coderの活用例が散見されるようになった。また同じ統計分析といっても，テキスト型データを対象とする場合に特有の手続きもある。よって，計量テキスト分析およびKH Coderを用いた論文の執筆・査読チェックポイントを整理しておくことには意義があると考えられる。

本稿で扱う３つのこと

KH Coderを用いた計量テキスト分析の結果を論文に掲載するときに留意すべき点として，本稿では報告・明示するべき点と分析結果の記述を主に扱う。

第一に，使用したテキスト型データや分析方法について，論文中で報告しておくべきことを確認する。基本的には，データ収集や分析の手順について，第三者が結果を再現しうるだけの情報を提示することが，論文執筆者としてフェアな態度である。

もちろん，インタビュー記録やフィールドノートのような質的データについては，第三者がそのまま再現することは不可能である。その場合でも，Y. S. Lincoln & E. G. Guba (1985) が質的研究における信頼性の一部として取り上げた，次のような点は維持すべきである。すなわち，研究者がたどったデータ収集・分析の過程をほかの研究者が監査（audit）できねばならない。これは幅広くさまざまな立場の研究において求められることであろう。

第二に，たとえば共起ネットワークや対応分析といった計量テキスト分析の結果を用いるときには，そこから何を読み取れるのか，なぜ読み取れるのかを文章で記述せねばならない。この記述は，必要な情報を明示しながら，順を追って行なわなければ，読者からは飛躍しているように見えたり，説得力に欠けたりといったことが起こりうる。さらには，上述の意味での信頼性が損なわれることもあるだろう。そこで本稿では，筆者らが適切と考える記述の形を示す。そのままテンプレートとしてもお使いいただけるし，そうしない場合も，明示すべき情報や記載の順序といった点を汲み取ることで，適切な記述の助けとしていただけるだろう。

第三に，論文に掲載すべき分析結果の選び方ないしは分析の進め方について，筆者らの推奨する点を述べる。たとえば，共起ネットワーク10個を論文に掲載して，それを読者に見比べさせることには無理がある。人間の目で比較を行なうのではなく，統計の助けを借りる形で比較を行なうべきである。こうした選択について述べるとともに，その他の留意点を確認する。

ここで最後に，本稿では扱わないことにも触れておきたい。論文執筆においてもっとも大切なことはおそらく，当該の研究分野における新規性と重要性を有することであろう。同じことが，実質科学的な知見があるかどうかという言葉で表現されることもある。これについては各研究分野において，論文の執筆者・査読者の皆さまに各自でご検討いただく必要がある。私達に書けるのは，論文では常に新規性・重要性を明確に説明すべきだということまでである。本稿の目的は，あくまでKH Coderによる計量テキスト分析の結果を，適切に論文に用いるための留意点を提示することにある。

１．報告・明示すべき点

1.1. データ収集の過程

データ収集の過程を，原則的には読者が再現できる程度の詳細さで報告することで研究の透明性がたもたれ，ひいては研究の質が向上する。以下ではア

ンケート自由記述やインタビュー記録を扱う場合のチェックポイントを例として挙げているが，異なる種類のデータを扱う場合にも同様の情報を示しておきたい。

1.1.1. 回答者の集め方（サンプリング方法）

ランダム・サンプリングにもとづく質問紙調査の場合には，母集団やサンプリング方法について具体的に記載すればよい。

森田ほか（2012）は都市イメージの研究において，系統無作為抽出を行ない2,118票を回収した。また，そのうち自由記述欄に記入があった1,614票を分析対象としたことを報告している（表１）。

表１　データ収集方法：森田ほか（2012）より抜粋

調査日	配布：2008年11月中旬から下旬 回収：2008年12月14日（郵送期限）
対象地域	群馬県前橋市利根川左岸地域（人口約9.9万人，約4.2万世帯，84町丁・大字，面積約27.6km^2）
抽出	系統無作為抽出（世帯抽出），　抽出率約9.5%
対象者	対象地域の4,000世帯の構成員
調査方法	配布：調査員による戸別配布（ポスティング） 回収：郵送回収
調査内容	1) 個人・世帯属性 2) 居住地区の生活質の満足度「あなたの住んでいる地区は住みやすいと思いますか」（5段階評価） 3) 自由記述 「前橋の良い点，好きなところ，場所などを，短い言葉や文で，自由に書いてください．」
回収数	世帯：1,293世帯（回収率32.3%） 個人：2,118票回収（うち自由記述欄に記入のある1,614票を有効票とした）

このように，調査全体への回答者数とは別に，自由記述型の設問に回答した人数を示す必要がある。後述するクレンジングで「特になし」といった回答を除外した場合は，分析対象として残ったケース数も示すとよい。

さらに，何文字記入されたのかという自由記述の長さを報告すれば，収集したデータについてより詳しく読者に伝えられる。文字数の平均と標準偏差を示せば，回答者が記入した長さのばらつきも含めて表現できる。なお文字数ではなく語数を用いてもよいだろう。

1.1.2. 自由記述をうながす設問文

たとえば「ご意見があればご自由にお書きください」という設問と，「○○について不満な点があればご記入ください」という設問とでは，えられる回答はまったく異なる[2)]。設問文の影響は大きく，単語が１つ含まれるかどうかによって回答が大きく変わることもあるので，設問文は全体を１字１句そのまま示しておきたい。

川端・樋口（2003）は，普及初期のインターネットに対する社会意識の分析において，設問全文に加え，その配置を示している。

> インターネットなどの情報技術に対する意識を探るために用いるのは，「情報と生活に関する調査」留め置き票の Q2である。この質問のワーディングを以下に示す。
>
> Q2. 最近，「インターネット」や「情報技術（IT）」という言葉を耳にすることが増えたと思います。情報技術（IT）やインターネットについてあなたが考えるとき，どんなことが思い浮かびますか。思いつくものから順に３つまで，何でもご自由にお答え下さい。
>
> 〔中略〕なお，記述内容が他の質問項目に影響されることを極力避けるために，この項目は質問紙の先頭近くに配置された。（川端・樋口 2003）

1.1.3. インタビュー記録の取り扱い

インタビュー記録を分析する際は，回答者（インタビュイー）の発言だけを分析することもあれば，聞き手（インタビュアー）の発言を含めて分析することもある。このどちらにしたのかを記載することで，読者も分析結果を理解しやすくなる。なお，対話全体の流れや，質問と応答の関係を把握したい場合には，聞き手の発言も分析に含める。たとえば「最近

の職場で不安に感じることはありますか」という問いに,「人員が減って業務が増えていることです」という回答があったとしよう。この場合，聞き手の発言を含めることで「不安」をあらわす内容であることが明確になる。

山崎ほか（2021）は，聞き手の発言を除外し，回答者である幼児教育アドバイザーの発言に注目したことを明示している。

> インタビューでは，幼稚園の環境（人的，物的），カンファレンスでの議論内容と，そこでの指導内容，訪問した幼稚園の先生方の幼児教育アドバイザーの受け入れや受容に係る内容，教師の保育上の悩みなどについて質問した。〔中略〕本研究で分析対象としたデータは，インタビュアーとインタビュイーの会話の内容を逐次文字に起こし，インタビュアーの発言を除いたインタビュイーの発言のみとした。（山崎ほか 2021）

1.1.4. クレンジング

分析に先立ってクレンジングを行なうことがある。たとえば類義語を統一したり,「特になし」のような意味のない回答を除外したりといった処理である。インタビュー記録や SNS の投稿, アンケート自由記述のように，表記揺れや記号の多いテキストを分析する際には，クレンジングを行なうことも多いだろう。クレンジングを行なった場合には，その内容を報告することが透明性の確保につながる。

勝谷・岡・坂本（2018）は，大学生を対象とした「新型うつ」に関する研究において, 分析データに対するクレンジングの内容を示している。

> 分析の際，手書きによる誤字･脱字を修正し，判読できない記述，記述途中の文，記号，頻度が多いが一般的でそれ自体特に意味を持たない語（例:「する」,「ある」）を分析からはずした。また,「うつ」,「鬱」のように意味が同じで異なる表記の語を１つに統一して分析に用いた。（勝谷・岡・坂本 2018）

1.2. 語の取り出し方

計量テキスト分析では，分析のための準備として，テキストデータから語を取り出す処理を行なう。語の取り出し方が変われば，分析結果も大きく変化するので，この部分も開示すべきプロセスの１つである。とくに，使用した分析ツールのデフォルトから変更した箇所については，論文中に記載しておきたい。

1.2.1. 形態素解析

日本語テキストから自動的に語を取り出すために，KH Coder では形態素解析を行なっている。形態素解析とは，文章を意味のある最小単位「形態素」に分割する処理である。基本的に KH Coder ではこの「形態素」を抽出語ないし語と呼んでいる。

形態素解析ツールを KH Coder 標準の「ChaSen」から「MeCab」に変更した場合や，形態素解析用の辞書に手を加えた場合には，そのことを記載すると分析プロセスが明確になる。

星野（2022）は新型コロナウイルスに関する新聞社説を分析する際に，新語に対応するために，MeCab 用の辞書「mecab-ipadic-NEologd」を用いたことを報告している。

> 通常の MeCab では ipadic という MeCab に同梱された辞書を参照しているが，この ipadic に含まれている単語は多くなく，また，新しい単語に対応していない。COVID-19に関しては，「三密」や「コロナ禍」といった言葉が生み出されているが，ipadic ではこれらに対応することができないのである。そこで，本論文では，この問題に対応するために,「mecab-ipadic-NEologd」という辞書を用いる。（星野 2022）

1.2.2. 強制抽出語と除外語

強制抽出語とは，指定した文字列を必ず１つの語

として抽出するKH Coderの設定である。辞書に収録されていない新語や専門用語を抽出できる。また，そのままでは複数の語に分割されてしまう語を，1語として抽出できる。たとえば，そのままでは「グループ」と「ワーク」や「看護」と「師」に分かれてしまうところを，「グループワーク」「看護師」として抽出できる。形態素解析用の辞書を入れ替えるよりも，KH Coder上で強制抽出語の設定を行なう方が手軽で便利な場合が多い[3)]。一方で除外語とは，意味の乏しい語や分析目的に照らして不要と判断される語を分析対象から除く設定である。

これらの設定を行なった場合は，強制抽出語や除外語のリストを示せばよい。さらに，これらの設定には分析者が何を重要で，何を不要と見なすかという方針が色濃く反映される。したがって，その方針を説明すること，すなわち強制抽出・除外の理由を説明することが，読者のよりよい理解や評価につながるだろう。

齊藤・田子・飯島（2021）は，分析対象とする専門用語を確実に抽出するため，「生ごみ」「再利用」「エコライフ」などを強制抽出語とした旨を明示している（表2）。また，井熊・中村（2023）は，南極に関する授業の感想文を分析する際，あえて「南極」を除外したことを理由とともに示している[4)]。

表2　強制抽出語：齊藤・田子・飯島（2021）より抜粋

生ごみ	b,e	油	e	NPO	a
ロス	d	4R	a	アプリ	a
リユース	b,e	Reduce	a	市民団体	b,b
再利用	c,e	ボカシ	a	Refuse	a
推進員	c,e	リフューズ	a	食育	b,d
エコライフ	b,b	母材	b,e		

右肩のアルファベットは元の形態素の品詞を次のとおり意味する。
a：未知語，b：名詞，c：サ変名詞，d：固有名詞，e：接辞（接頭，接尾）

なお，南極クラスとしての感想文の性質上，「南極」という抽出語が多く出現することが想定された。このような一般的な語は，共起ネットワーク分析を行う際の解釈性を低下させるため，「南極」を分析対象から除外する対応を行った。（井熊・中村 2023）

1.2.3. 総抽出語数・異なり語数

テキストから語を取り出す処理が完了したら，総抽出語数と異なり語数を示すことで，データの規模と語の多様性とを読者に伝えられる。総抽出語数とはテキストに含まれるすべての語の数を合計した，延べ数である。異なり語数とは，その中から重複を除いた語の種類の数である。これらの数値は，語の多様性や表現の幅を把握する手がかりとなるので，紙幅に厳しい制約がなければ記載しておくとよいだろう。

清水・荒井（2023）は，高大連携の実践を対象とした研究において，これらの数値に加えて文の数を示し，分析データの規模感や多様性を伝えている。

本研究の対象者のうち，175人から自由記述の回答が得られた。得られた回答データについて，総文数は238，総抽出語数は4,612，異なり語数は740であった。（清水・荒井 2023）

1.2.4. 品詞の選定基準

たとえば形容詞・副詞だけを分析に含めることで，調査対象者の主観的な認識に注目できるかもしれない。分析の目的に応じて，分析に含める品詞は変更できる。

分析に使用する品詞を，KH Coderの標準設定から変更した場合は，変更内容とあわせて，変更の理由を説明しておくと読者に分かりやすい。

福島・今井（2022）は，教師の用いる語彙の分析において，品詞の選定方針を明示している。

このうち，単語が単独で意味をもっていない助詞，感動詞，助動詞や，その学級特有の固有名詞を集計から除外した上で，5回以上使用した名詞，動詞，形容詞，形容動詞，副詞を教師使用語彙とし，これは67語であった。本研究では，

この教師使用語彙を対象に，教師の発話に用いられる語彙の特徴を検討する。(福島・今井 2020)

1.2.5. 抽出語と出現頻度のリスト

紙幅に余裕があれば，分析に使用した語の一覧を，それぞれの語の出現頻度とともに示すのが望ましいだろう。たとえば共起ネットワークや対応分析のような分析方法は，語を網羅的に示すという目的には必ずしも適していない。データ中に多く出現した語を網羅的に示し，重要な語が抜けていないか，おかしな語が混入していないかを確認できるようにするには，抽出語の一覧が適している。

狩野・三浦（2019）は，インタビュー記録を分析した研究において，抽出語とその出現頻度を上位150語までリスト形式で明示している（表３）。また，古口ほか（2022）は，出現頻度が高い語ほど語彙数が少なく，低頻度語ほど語彙数が多くなるという特性を利用して，紙幅を抑えつつ一覧を示すという工夫を行なっている（表４）。

２．分析結果の記述

2.1. 共起ネットワークの示し方

共起ネットワークとは，データ中でともに出現することが多い語同士を線で結び，ネットワークを描く可視化の方法である。テキスト中に多く出現する主題や，主題同士のつながり方を把握するのに適している。この図から読み取れることを，読者に伝わ

表４　抽出語リストの別例：古口ほか（2022）

出現回数	単語
68	対策
63	新型コロナ
62	クラスター
61	感染
24	ウイルス
20	定義
19	用語
18	ゼミ
17	続く
16	解説
14	緊急事態宣言，調査
13	人，接触
12	情報，専門
11	時間，例
10	発信
9	お願い，月，状況，地域
8	割，行動，質問，投稿，特定
7	回答，考え方，実効再生産数，対象，読む，必要，平均
6	拡大，行う，制限，西浦，説明，多い，注意，動画，班，保健所
5	NIPH, チーム，データ，押谷，解除，皆様，基本再生産数，研究，今後，思う，示す，多く，分かる，方々，流行
4	たくさん，リプライ，画像，外出，機能，距離，協力，結果，検査，言う，考える，参考，山，資料，社会，集団，正しい，生活，知る，都道府県，濃厚，倍加，発生，分析，北海道大学
3	ありがとう，インフルエンザ，グラフ，ヒト，ページ，リスク，リンク，アール，移動，医療，疫学，応じる，加える，該当，企業，嬉しい，経路，見る，弧，指摘，事業，修正，出す，詳しい，詳細，状態，申し訳，診断，全国，措置，総合，増える，増加，調べる，頂ける，訂正，努める，東北大学，日本，倍，変更，密，理解，理由

表３　上位150語のリスト：狩野・三浦（2019）より抜粋

抽出語	回数	抽出語	回数	抽出語	回数	抽出語	回数	抽出語	回数
思う	689	言う	73	サイン	42	希望	30	見直し	22
本人	597	地域生活	72	訪問	42	症状	30	相談支援専門員	21
プラン	319	家族	68	様子	42	薬	30	提案	21
クライシス・プラン	229	共有	68	方法	41	ケア会議	29	服薬	21
支援	228	多い	64	感じ	40	利用	29	立場	21
状態	82	周り	44	中心	31	調整	23	得る	17
難しい	80	修正	44	判断	31	特に	23	認識	17
考える	79	立てる	44	会議	30	把握	23	移行	16
計画	77	機関	43	管理	30	福祉	23	教える	16
確認	76	前	43	危機	30	気持ち	22	緊急	16

りやすい文章で記述するためのチェックポイントを以下に示す。なお共起ネットワークに限らず，クラスター分析・多次元尺度法（MDS）・自己組織化マップなどの分析にも同様のポイントが当てはまる。

2.1.1. 集計単位

同じ文のなかに２語とも出現していれば共起していると見なすのか，それとも同じ段落のなかに出現していれば共起とみなすのか。こうした文・段落・Excel のセルといった単位のことを KH Coder では「集計単位」と呼んでいる。KH Coder は，ユーザーが選択した集計単位の中に共起する語を線で結び，共起ネットワークを作成する。

たとえばアンケート自由記述では，１つの回答のなかに出現していた語を，共起していたと見なして線で結ぶのが一般的である。しかし，集計単位を文にすることで，同じ文のなかに出現していた語だけを線で結ぶこともできる。前の文と後の文とで話題が変わることが多い場合や，直接的な意味の関連をもつ語同士に注目したい場合には，集計単位を文にすることも考えられる。

このように集計単位の設定次第で，共起ネットワークの読み取り方も変わるので，どの単位を選んだのかを記載しておきたい。また何らかの理由で能動的に単位を選択した場合は，その理由を書いておくと分かりやすい。

川本・渡辺・日高（2018）は，集計単位として「段落」を選んだことを明示している。さらにクレンジングの一環として，１つの回答内に記載された内容であっても，主題が明らかに異なる場合には改行を加えて別の段落としている。主題ごとのまとまりをたもった分析結果になるようにとの配慮であろう。

> （１）前処理：〔中略〕同一セル内で明らかに意図が異なる内容の場合は，改行して別段落とした。〔中略〕
> （２）クラスター分析：テキストデータを内的要因と外的要因の２つに分けて，それぞれで KH Coder を用いて，文書クラスター分析を行った。また，段落ごとを分析単位とし，Ward 法と Jaccard 距離を用いた。（川本・渡辺・日高 2018）

2.1.2. 語のグループに名前を付ける

共起ネットワークをいくつかの部分に区切り，各部分に集まった語のグループに名前を付けるというのが，一般的な解釈の進め方である。語のグループに対して，それらの語がデータ中で表現している主題の名前を付けていくことで，データ中の主題が明らかになっていく。

図の提示においては，共起ネットワークの区切りや，各グループに付けた名前を書き込むと，分かりやすい図になる。また文中で言及するために，各グループに番号を付すのも一手である。図を示すスペースに余裕があれば，図中の語を含む実際の文章表現を，吹き出し等を使って書き込むとリアリティが増すだろう。

図を記述する文章では，各グループ名について，そのように名付けた理由を読者が納得できるようにしなければならない。グループ内の主な語をいくつか挙げた上で，それらの語がデータ中で利用されていた文脈を説明するのが一般的である。その際には，KH Coder の KWIC: Key Words In Context 機能を用いて，それらの語を含むデータ中の原文を示すことが有効である。

奥村・香月（2024）は，共起ネットワーク中のグループに命名する際，各グループを特徴づける代表語と，その語が実際に用いられた原文を提示している（表５）。さらにそれぞれの原文について，回答者の属性もあわせて記載することで，命名の根拠となる記述が一部の回答者に偏っていないことを示している。

なお同じ言葉でも，使われる文脈によってその意味が変わることはめずらしくない。しかし「ほかの特定の語と共起していること」という条件が加われば，文脈や意味がはっきりする場合も多い。多義的に用いられる語の場合には，共起ネットワークにお

表５　KWIC 機能で検索した原文：奥村・香月（2024）より抜粋

テーマ	抽出語（コード数）	KWIC コンコーダンスに基づく代表例
支えてくれる人	家族（22）	安らぎや**家族**・ペット等の存在（51 歳 男性） 支えてくれる**家族**には大変感謝しています（48 歳 女性） 「そのままで大丈夫」というメッセージをくれる**家族**・友人の存在はとても助かります（43 歳 女性）
	友人（16）	自分の羅針盤となる支援者や**友人**の存在（57 歳 男性） 家族など，**友人**など，ペットなどが，私の人生にとても大事です（27 歳 男性） デイケアで**友人**を作ること（57 歳 男性）
	デイケア（20）	日頃思う現実に落ち込まないように，**デイケア**に通所して明るく生きること（50 歳 女性） **デイケア**で同じような病気を持つ人の話が聞けてすごく嬉しい（35 歳 女性） **デイケア**でプログラムを受けたり，皆と交流することが役立っています（46 歳 女性）
	スタッフ（10）	**スタッフ**さん達や先生の助言など（50 歳 女性） 友人やデイケアの**スタッフ**などとの関わり（26 歳 男性） 「自分らしい生き方」を肯定してくれる**スタッフ**さんや家族・友人がいてくれると嬉しい（37 歳 女性）

いて線でつながっている語を挙げて説明するのもよいだろう。すなわち「語Ａや語Ｂと共起していることから，○○の意味で使われていたことが分かる」といった説明を加えるのも一案である。

2.1.3.　グループ名についての追加的な工夫

グループ名が適切であるかどうかについて，もしも査読者から疑問が呈された場合には，さらに次のような追加的な工夫も考えられる。

１つの工夫として，それぞれのグループ名に合致する記述が何件あったかを示すことが考えられる。言葉は多義的に使われるものであるから，グループ名と合致する意味で使われた回数と，異なる意味で使われた回数を数えて示すということである。これによって，グループ名に合致する記述の割合を把握できるので，過剰な解釈や，限定的すぎる主張を避けることができる。

川上ほか（2022）は，グループを特徴づける代表語を含む記述の数と，そのなかでグループ名に合致する記述の数を報告している（表６）。

もう１つの工夫として，グループ名を検討するための体制を整え，それを報告することも考えられる。たとえば複数名の研究者でグループ名を検討したり，調査対象者に命名を確認してもらったり，専門的知見を有するスーパーバイザーの監修を受けるといったことが考えられる。

表６　グループ名に合致する記述件数：川上ほか（2022）より抜粋

種別	カテゴリー		主な共起語	記述件数	主な自由記述（太字は主な共起語）
利点	H1	実際の様子を見学できる	様子，見学，見る	10/14	「実際に診察の**様子**を**見学**できた」「診察の**様子**を**見学**することができ，どのようにして認知症患者と関わっているのか学ぶことができた」「援助やリハビリなど様々な場面を**見る**ことができた」
	H2	実際の現場で学ぶことができる	現場，実習，学べる	11/16	「やはり**現場**でしか**学べ**ないことを**学べ**た点」「実際の**現場**の空気の中で**実習**をやりきったことが自信につながった」
	H3	手技を練習できる	演習，手技	4/8	「指導を受けながら実際に看護**手技**をさせてもらえた」「**演習**で学んだ**手技**を実習で活かすことができた」
	H4	患者と向き合う時間が長い	受持つ，長い，多い	7/7	「患者を**受持ち**，**長い**時間関わって生活史のことをたくさん知ることができた」「実習時間が**長く**，患者さんと向き合える時間が**多かっ**た」
	H5	病院に行ける	病院，病棟，行く，行ける	8/9	「実際に**病院**に**行っ**て学べた」「**病院**に**行け**たこと自体がよかった」「半日でも**病棟**に**行く**ことができた」
	H6	体験を看護計画に活かすことができる	体験，アセスメント，計画，実施	10/13	「受け持ち患者の手伝いを**実施**して，対象者にとって何が必要なことか実際に**体験**できた」「患者本人の思いや気持を考慮して，それを**アセスメント**や**計画**に反映することができたところ」
	H16	実習期間が短く物足りない	病院，期間，物足りない	4/6	「**病院**実習へ行かせてもらえたのはありがたいが，**病院**実習の**期間**には**物足りな**さを感じる」「**病院**実習の**期間**が短く，計画の立案から実施，評価の段階まで到達することができなかった」

注）記述件数は，左：主な共起語を含んでおり，かつ，カテゴリーに示した意味で解釈できる記述件数，右：主な共起語を含むすべての記述件数

石徹白・山田（2024）は命名の際に，調査対象者への確認（メンバーチェッキング）を行なうとともに，がん看護に携わる専門家によるスーパーバイズを受けたことを報告している[5)]。

> 調査対象者に対して抽出された語からクラスター名の解釈において対象者の ACP を行う上でリンクナースが感じる障壁が正しく反映されているかメンバーチェッキングを行った。またがん看護に携わる専門家によるスーパーバイズを受けた。（石徹白・山田 2024）

2.1.4. グループの柔軟な分割

KH Coder には共起ネットワークを自動的にグループ分けする機能があるが，必ずしも結果をそのまま利用する必要はない。KH Coder が分割した 1 つのグループ内に複数の主題がある場合は，それらを切り分けて複数の名前を付ければよい。また KH Coder が複数の異なるグループに分割したとしても，同じ主題に関連していると判断される場合は，同じ名前を付ければよい（図 1）。機械的に分割された結果を絶対視せずに，語の意味と文脈にもとづいて柔軟に命名されたい。

2.2. 対応分析の示し方

たとえばアンケートの場合「20代女性」「20代男性」「30代女性」…といったカテゴリー分類によって，自由記述がどう変わるかを見るために，対応分析を利用できる。対応分析の結果を示し，読み取れることを記述する際のチェックポイントを以下に示す。

2.2.1. カテゴリーの数と寄与率

たとえば「20代女性」「20代男性」…のようなカテゴリーの数は，多くとも20程度までに抑えた方がよいだろう。カテゴリーが多すぎると分析結果が複雑になって解釈が難しくなり，成分の寄与率も小さくなる傾向があるからだ。仮に10パーセントを下回る寄与率の成分ばかり多数抽出された場合，どの成分に注目するのかという問題も生じる。カテゴリー数を絞り，横軸と縦軸の寄与率がおおむね50% を大きく下回らなければ，KH Coder が作成する 2 次元のプロットを用いて問題ないだろう。カテゴリーと抽出語との対応関係を，50% 程度は表現できているということだからだ。

Sakai et al.（2024）は学術論文を対象としたトピックの分析において，出版年（1990年～2022年）をそのまま使用せず，「1995-1999」のような 7 つのカテゴリーに分類している。これによって解釈しやすく，寄与率も高い対応分析の結果を示している。

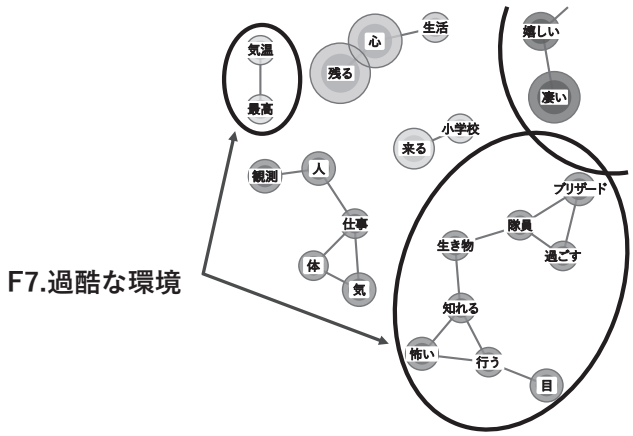

図 1　複数のグループに同じ名前を：井熊・中村（2023）より抜粋

2.2.2. 各カテゴリーに特徴的な語

対応分析のプロットでは，原点（0, 0）の近くには平均的な出現傾向の抽出語が集まる。もしカテゴリー名が原点の近くに付置された場合，そのカテゴリーには平均的な語が含まれていた，すなわちデータ全体と同じような語が含まれていたことが分かる。逆に，原点から遠く離れた抽出語やカテゴリー名には，平均から離れた何らかの特徴があったと見なせる。

したがって，対応分析のプロットを解釈する際には，原点の遠くに離れた抽出語から順に確認していくとよい。そして，原点から見て「20代女性」の方向に遠く離れた語は，「20代女性」の回答に特に多かった，「20代女性」に特徴的であったと解釈できる。

対応分析は，このように原点からの方向と距離にもとづいて解釈する。仮に２つの語が「20代女性」から同じ距離にあったとしても，その語が原点側にあるのと，原点から遠い側にあるのとでは，特徴的である度合いが異なる。よって「20代女性」のようなカテゴリー名と，抽出語との距離を見て「近い」「遠い」と解釈することは厳密には誤りである（樋口・中村・周 2022）。

安藤（2023）は，インタビューデータの分析において，分析結果の中で原点から離れて配置された語に注目し，カテゴリーによる語の変化を解釈している。

> なお対応分析では，原点からの距離が遠くにある語ほど何らかの強い特徴があると解釈〔中略〕される。社会性・組織効力感高群に特徴的な語として，遠く離れて布置されている「話す」に加えて，「聴く」「管理」（者・室）「評価」（システム・制度）が見られた。（安藤 2023）

以上のように特徴語を確認したうえで，それぞれの特徴語がデータ中で使用されていた文脈を説明する必要がある。データ中の原文を提示しながら説明すると良いだろう。これは共起ネットワークやその他の分析とも共通することだが，統計だけを見て判断することは現代の技術水準では危険である。必ず原文に戻って確認しながら解釈を行なうのが安全であり，論文中にもなるべく原文を引用したい。こうした考え方から KH Coder の共起ネットワークや対応分析では，プロット上の語をクリックするだけで，その語の前後の文脈を閲覧できるようにした。

2.2.3. それ以外に読み取りうること

カテゴリー名と抽出語との距離は解釈できないが，カテゴリー名同士の距離や，抽出語同士の距離は解釈できる。２つのカテゴリー名が近くに付置されていた場合，それらのカテゴリーは語の出現傾向が似通っていたと解釈できる。ただしカテゴリーの数が３つまでの場合，各カテゴリー名は離れて付置される場合が多い。カテゴリー数が４つ以上の場合は，どのカテゴリーとどのカテゴリーが近くに付置されているのか，すなわち内容が似通っていたのかを読み取るのもよいだろう。

次に，軸ごとに解釈を行なうことで，新たな気づきをえられることもある。縦軸を無視して，横軸だけに注目し，どんな抽出語やカテゴリー名が右側の端にあるのか，左側の端にあるのかを見るような形である。ここでは，上部か下部かは関係なしに，とにかく右側／左側にある抽出語やカテゴリー名を見ればよい。このように軸ごとに解釈をすることで，横軸にあらわれた変化と，縦軸にあらわれた変化とを別々に把握して，軸に名前を付けられるかもしれない。ただし，必ずしも軸があらわす変化の内容を明らかに読み取れる場合ばかりではないので，無理に軸に名前を付けようとする必要はない。

2.3. コーディング結果の示し方

分析者が注目したいと考えるコンセプトを柔軟に取り出す方法としてコーディングがある。「語１・語２・語３のいずれかが出現していれば，コンセプトＡが出現していたと見なす」といった指定を，分析目的に応じて自由に行なえる。KH Coder ではこの

指定のことをコーディングルールと呼んでいる。コンセプトの出現頻度をカテゴリーごとに集計・比較したクロス集計や，クロス集計を可視化したバブルプロット・折れ線グラフなどが多用されている。

コーディング結果を示す際には，コーディングルールを提示することで，分析者がどのようにコンセプトを取り出したかを示す必要がある。コーディングルールは，質問紙調査における設問文と同様に，自由かつ創造的に作成することが推奨される。そして設問文と同様に，公開することが原則である。コーディングルールの公開は，検証を可能にするだけでなく，他の研究者が同じコンセプトを用いて研究を継承し発展させることにもつながる。

高（2015）は，現代的レイシズムに関する研究において，独自のコーディングルールを作成し，その内容を明示している。このコーディングルールは後続の研究においても，データに合わせて修正した上で活用されている（田村・田村 2016; 曹 2017a, 2017b）。

紙幅の都合により，コーディングルールを本文中に記載できない場合には，別表やオンライン補足資料として全体を公開することが望ましい。その際には，仮に全体を掲載できなくとも，要約版を論文に示しておくとよいだろう。

雲財（2022）は，分析に使用したコーディングルールを Web 上で公開するとともに，分析データもあわせて提供している。

> 本研究で使用したデータセット，出力，コーディングは，次の OSF のリンクにて公開している。〔後略〕（雲財 2022）

3．分析結果の選択と留意点

3.1. 共起ネットワークを多数並べない

たとえば「20代女性」「20代男性」「30代女性」…のようなカテゴリー別に共起ネットワークを多数作成し，それをすべて示すのは避けるべきである。多くの共起ネットワークを目で見比べて，抽出語やテーマの重なりと差異を見分けることは困難である。

それに対して対応分析であれば，各カテゴリーの特徴や，カテゴリー間の類似性を１つのプロットで表現できる。よって対応分析を行なった上で，必要に応じて，特に注目したい１つか２つのカテゴリーの共起ネットワークを示すとよいだろう。

必ずしも対応分析でなくとも，各カテゴリーの特徴語をまとめた表や，カテゴリー名と抽出語を線で結ぶ形の共起ネットワークでもよい。重要なのは目で比較するのではなく，計算によって特徴語をリストアップするといった形で，統計の力を借りた比較を行なうことである。

3.2. 論文でワードクラウドを使用しない

ワードクラウドは，データ中に多く出現する語ほど大きく表示しつつ，語をランダムに配置した図である。「なんだかおもしろそう」と思わせるデザイン性がありインパクトも強い。表示するのは語だけなのでスペース効率もよく，限られたスペースでも提示しやすい。こうした理由から，ワードクラウドを印象的で興味深いグラフィックとしてプレゼンテーションに活用することは有効であろう。

一方で論文における利用には疑問が残る。出現回数が最上位の何語かについては明確に把握できるが，それ以外については，見る人によって目に入る語が異なるだろう。人によって異なる印象を受けることは，アカデミックな場面では好ましいとはいえない。論文では，出現回数を正確に把握できて，出現回数が多い順に語を閲覧できるリストの方が読み取りやすい（表３・表４）。

またランダムに語が配置されていると，１語１語を別々に見ざるをえない。それよりも，語と語のつながりを読み取れる点で，共起ネットワークの方が図として優れているだろう。共起ネットワークをもとに，多く出現した語ほど大きく表示する「文脈クラウド」機能も KH Coder に備えたので（図２），こちらも利用されたい。

図２　文脈クラウドの例

3.3. KH Coder のバージョン番号

KH Coder は継続的に更新され，機能の改良が行なわれている。その際, 同じ操作や設定であれば, 同じ分析結果がえられるように細心の注意を払っている。KH Coder をアップデートしても，分析結果が勝手に変わってしまうということはまず起らないだろう。

しかし，適切とはいえないオプション設定が可能になってしまっていた場合，その設定を選べないように仕様を変更するといったことはある[6)]。このため，使用した KH Coder のバージョン番号を論文中に正確に記載することが望ましい。これにより他の研究者による追試や検証が容易になる。なお現行の KH Coder のバージョン番号のフォーマットは「Version 3.** official-package」となっている。

3.4. KH Coder に固有の用語

KH Coder を用いた分析結果を論文に用いる際には，KH Coder に固有の用語（e.g., 外部変数，未知語，名詞 B，名詞 C）を使いたくなる場面があるだろう。こうした用語を論文中で用いる場合には，初出時にその意味や役割を簡潔に補足することが推奨される。これによって KH Coder になじみのない読者や異分野の読者にも理解しやすくなるだろう（表７）。

表７　KH Coder 固有用語と注釈の例

固有用語	注釈
外部変数	たとえば性別・年代カテゴリーのようなテキスト以外の変数
総抽出語数	抽出語の延べ数
異なり語数	抽出語の種類数
強制抽出語・タグ	強制的に１つの語として抽出するよう指定した語
未知語	語を取り出すための辞書に掲載がなかった未知の語
名詞 B	「もの」のような，平仮名のみの名詞
名詞 C	「人」のような，漢字１文字の名詞

おわりに

本稿では，計量テキスト分析の実践において，とくに KH Coder を用いた研究論文を執筆・査読する際のチェックポイントを示した。KH Coder を念頭において記載しているものの，ほかの分析ツールを用いる場合にも，おおむね同じチェックポイントが

当てはまるだろう。

本稿で示したチェックポイントは，分析者自身が分析結果を論文にする際のガイドラインとして活用できるだけでなく，査読者が論文を評価する際のチェックリストとしても有効であろう。とくに近年では，社会学だけでなく，心理学・教育学・看護学など，計量テキスト分析の活用が広がる一方で，分析の信頼性や記述の妥当性についての評価基準が共有されにくいという課題がある。本稿がこの課題に対応するための一助となれば幸いである。

付記

本稿は，2024年9月のKH Coder公式セミナー・リサーチ実践編にて口頭で報告した内容に，加筆・修正を加えたものである。

注

1）KH Coderは2023年12月より株式会社SCREENアドバンストシステムソリューションズより産学連携製品として販売されることになり，無償提供を終了した。

いくつかの制限はあるものの無料版「Startingエディション」のほか，先生方向けの授業用ライセンスもあり，無料でKH Coderを実習にお使いいただける。またアカデミック割引や，さらに割安の「ゼミ＆研究室パック」も準備されているので，ご活用いただけると幸いである。

2001年から20年ほどのあいだ本稿執筆者の1人である樋口耕一が，開発・メンテナンス・サポートなどのすべてを担うことで，KH Coderの無償提供を実現してきた。しかしこの体制は限界で，継続できなかった。樋口に体調不安が生じことや，株式会社SCREENアドバンストシステムソリューションズがKH Coder開発の趣旨に賛同したことから，産学連携によって同社が協力する運びとなった。これによってKH Coderは，従来よりも信頼性と継続性に優れたソフトウェア製品となった。なお同社所属の中村康則と周景龍が本稿の執筆にも携わっている。

2）「ご意見があればご自由にお書きください」という設問文の場合，あまりに多様な内容が書き込まれ，結果として計量的な分析が困難になる場合が少なくない。

3）分析に用いるのは出現回数が上位の100語から200語程度というケースも多い。共起ネットワークや対応分析に用いる上位100語程度が正確に抽出されればよい場合，辞書を入れ替えるよりも，いくつかの用語を強制抽出語として指定する方が簡便である。一方で，より多くの語を分析に用いたり，非常に多くの新語や専門用語がデータに含まれる場合には，辞書を入れ替えるとよいだろう。星野（2022）の用いた「mecab-ipadic-NEologd」のほか，KH CoderのWebページで医療用語の辞書を公開している。

4）共起ネットワークを作成する際に，どの回答中にも出現するような一般的な語を含めると，ほかの多くの語と線でつながってしまって図全体の解釈が難しくなることがある。

5）この研究は，階層的クラスター分析におけるクラスターの命名に際して，専門家のスーパーバイズを受けたものである。クラスターは，共起ネットワークにおけるグループ分けと同様の機能を果たすため，本文中では両者を関連付けて引用した。

6）共起ネットワークの「最少スパニングツリーのみを表示」オプションは，語と語のネットワークにおける使用を念頭において備えた機能である。カテゴリー名と語のネットワークにおいてこのオプションの利用は適さないが，選択可能となっていたので，KH Coder Version 3.Beta.07以降では選択不可とした。

参考文献

安藤美華代（2023）．中間管理職からみたキャリアのある新任医療従事者のメンタルヘルス支援に関する探索的検討．文化共生学研究，23，1-11.

曺慶鎬（2017a）．“Yahoo! ニュース”の計量テキスト分析：中国人に関するコメントを中心に．駒澤社会学研究 Komazawa journal of sociology, (49), 115-135.

曺慶鎬（2017b）．インターネット上におけるコリアンに対するレイシズムの対策と効果――“Yahoo! ニュース”のコメントデータの計量テキスト分析．

応用社会学研究，59，113-27.

福島（浦田）貴子・今井亜湖（2022）．日本語指導が必要な児童が授業に参加するために必要な語彙の検討．日本教育工学会論文誌，46(Suppl.)，77-80.

樋口耕一(2004)．テキスト型データの計量的分析―２つのアプローチの峻別と統合―．理論と方法，19(1)，101-115.

樋口耕一(2020)．社会調査のための計量テキスト分析――内容分析の継承と発展を目指して―― 第２版．ナカニシヤ出版.

樋口耕一・中村康則・周景龍（2022）．動かして学ぶ！はじめてのテキストマイニング：フリー・ソフトウェアを用いた自由記述の計量テキスト分析．ナカニシヤ出版.

星野雄介(2022)．自然言語処理技術を用いた新型コロナウイルスに関する新聞社説の予備的分析：新聞社ごとの違いと研究の展望．武蔵野大学経営研究所紀要，(5)，113-148.

井熊英治・中村康則（2023）．キャリア教育「南極クラス」における対面授業と教師支援型遠隔授業の違い．コンピュータ & エデュケーション，55，68-73.

石徹白しのぶ・山田忍（2024）．がん患者へのアドバンス・ケア・プランニングを行う上でリンクナースが感じる障壁の検討．日本看護研究学会雑誌，47(2)，175-187.

狩野俊介・三浦剛（2019）．精神障害者の地域生活支援におけるクライシス・プランの作成と活用のあり方に関する研究――精神科病院と相談支援事業所のソーシャルワーカーへのインタビュー調査――．社会福祉学，60(1)，47-62.

勝谷紀子・岡隆・坂本真士（2018）．大学生を対象とした「新型うつ」のしろうと理論の検討．心理学研究，89(3)，316-322.

川端亮(2021)．ライフヒストリーにおける言葉のあいまいさの活用　内田諭・大賀哲・中藤哲也（編）知を再構築する 異分野融合研究のためのテキストマイニング，ひつじ書房.

川端亮・樋口耕一（2003）．インターネットに対する人々の意識：自由回答の分析から．大阪大学大学院人間科学研究科紀要，29，162-181.

川上祐子・中村康則・松尾綾子・江本厚子（2022）．COVID-19下の老年看護学実習の形態別による到達度評価と課題の検討．日本教育工学会論文誌，46(4)，667-678.

川本弥希・渡辺雄貴・日高一義（2018）．高等教育における学習者のラーニングエクスペリエンスの形成に影響を与える要因．日本教育工学会論文誌，41(4)，363-374.

古口凌太郎・石橋由基・田谷元・安保悠里子・伊豆藏栞・奈良由美子・堀口逸子（2022）．新型コロナ対策専門家 Twitter アカウントにおけるツイートの特徴と情報発信内容．リスク学研究, 31(3), 219-229.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications.

森田哲夫・入澤覚・長塩彩夏・野村和広・塚田伸也・大塚裕子・杉田浩（2012）．自由記述データを用いたテキストマイニングによる都市のイメージ分析．土木学会論文集，68(5)，315-323.

奥村智志・香月富士日（2024）．精神保健医療ユーザーが知覚するリカバリーに効果的な要素．日本精神保健看護学会誌，33(1)，10-18.

齊藤由倫・田子博・飯島明宏（2021）．計量テキスト分析を用いた自治体の生活系ごみ減量化啓発事業の分類．廃棄物資源循環学会論文誌，32，13-127.

Sakai, K., Nagata, T., Mori, T., Inoue, S., Fujiwara, H., Odagami, K., Adi, N. P., Takemichi, M. & Mori, K. (2024). Research topics in occupational medicine, 1990–2022: A text-mining-applied bibliometric study. *Scandinavian journal of work, environment & health*, 50(7), 567-576.

清水優菜・荒井英治郎（2023）．総合的な探究の時間における高大連携の効果の検討．日本教育工学会論文誌，47(1)，47-61.

高史明(2015)．レイシズムを解剖する――在日コリアンへの偏見とインターネット．勁草書房.

田村貴紀・田村大有（2016）．路上の身体・ネットの情動――3.11後の新しい社会運動：反原発，反差別，そして SEALDs．青灯社.

雲財寛（2022）．『理科教育学研究』はどのような論文を掲載してきたのか―テキストマイニングを用いて―．理科教育学研究，63(1)，225-232.

山崎晃・松井剛太・越中康治・濱田祥子・東和子（2021）．

幼児教育アドバイザーの幼稚園への働きかけの変
容を捉える．子ども学論集，７，15-26.

Checkpoints for Writing and Reviewing Academic Papers Using Quantitative Text Analysis and KH Coder:
Reporting Results with Elements Requiring Explicit Presentation

NAKAMURA Yasunori[i], SHU Keiryu[ii], HIGUCHI Koichi[iii]

Abstract : Quantitative text analysis refers to a methodological approach in which textual data are statistically processed for the purpose of conducting content analysis. KH Coder, a software tool developed by the authors for such analyses, has been used in more than 8,000 academic studies, particularly in the social sciences and humanities. In this paper, we outline a set of checkpoints designed to ensure the appropriate writing of academic manuscripts based on results obtained through quantitative text analysis using KH Coder. Each checkpoint is illustrated with examples of effective practices, making reference to actual research cases whenever possible. These checkpoints are also expected to provide useful guidance for peer review of manuscripts that apply quantitative text analysis.

Keywords : quantitative text analysis, text mining, survey of studies using KH Coder, peer review

i Part-time Lecturer, Department of Human Health Sciences, Graduate School of Medicine, Kyoto University; Corporate Officer, SCREEN Advanced System Solutions Co., Ltd.
ii Manager, Development Department, SCREEN Advanced System Solutions Co., Ltd.
iii Professor, College of Social Sciences, Ritsumeikan University